Introduction to NLP

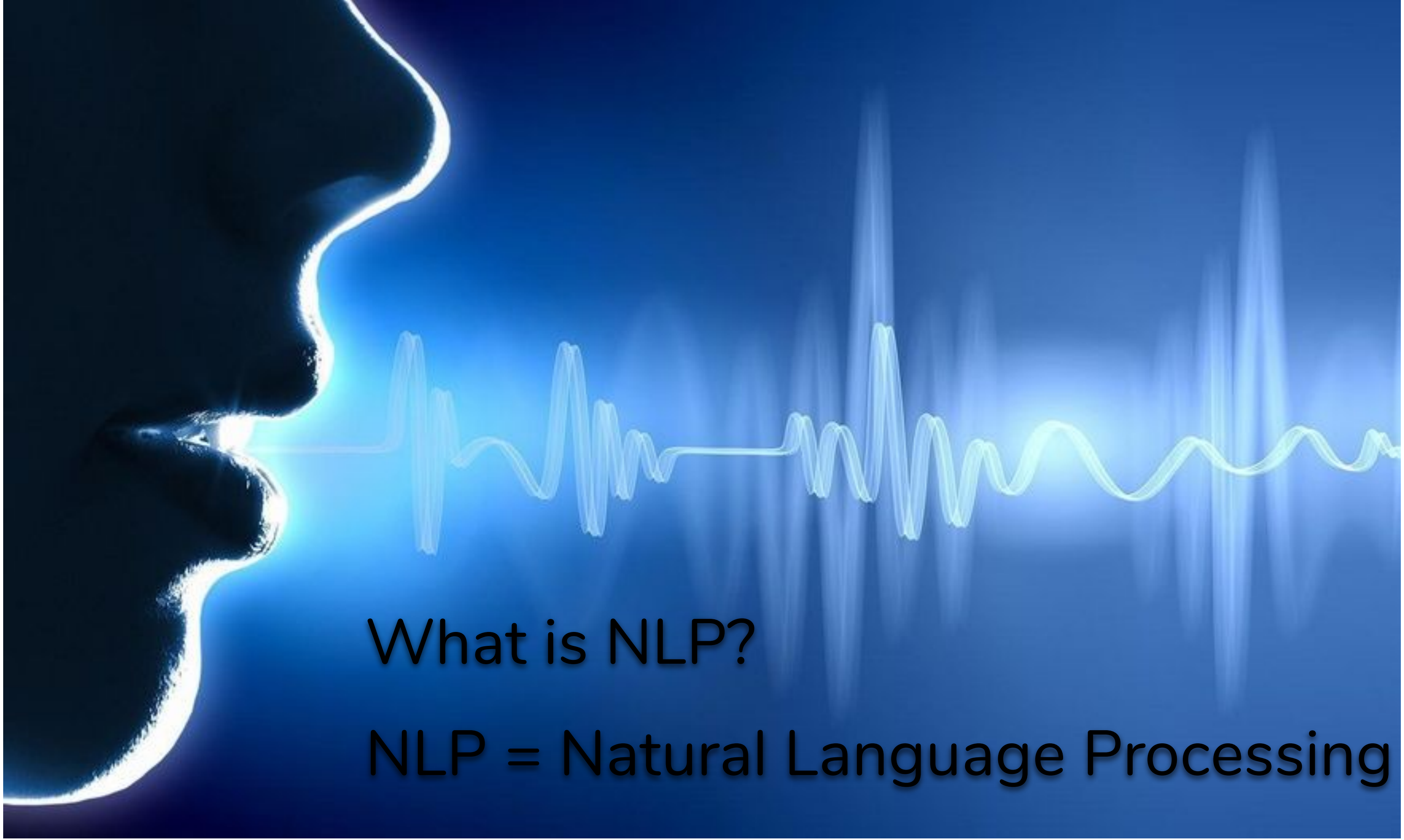
Adela Muresan
Head of Machine Learning @ Banca Transilvania

Can you understand a text?

Can you say what it is all about?

Of course!
You can!

Me too!



What is NLP?

NLP = Natural Language Processing

Processing ....

Parsing, sentence segmentation and part-of-speech tagging - understand the grammatical construction of sentences

Entity Extraction - identify different entities like geographic locations, first, last names, company names, addresses, phone numbers, company names, etc.

Machine Translation - automatic translation

Automatic Summarization - ex: a short summary of a lengthy academic article

Co-reference resolution - which words/phrases are used to refer to the same objects

Sentiment analysis - a text is positive/negative/neutral

Processing ….

Natural language generation - Convert information from computer databases or semantic intents into readable human language

Natural language understanding - involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression

Optical character recognition (OCR) - Given an image representing printed text, determine the corresponding text.

Topic segmentation and recognition - Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment

Word sense disambiguation - select the meaning which makes the most sense in context.

Chatbots - interact with computers through language

Where could you apply it?
Usages in Fintech World

Digital Financial Coach/Advisor

In finance, transactional bots can be used to offer users finance coaching/advising services
An example of this application has been deployed by Sun Life(Canadian life insurance company) which created a virtual assistant Ella, to help users for Benefits and Pension by allowing them to stay on top of their insurance plans. The assistant sends users reminders based on user data like "Wellness benefits about to expire" or "Your child will be off benefits soon."

Automated Claims Process

Totally useful in the insurance world.

Transaction search & visualization

Bank of America uses such a bot (called Erica) as a digital financial assistant for their clients base. The AI-powered bot was quickly adopted —1 million users in threee

“For instance, a bot might tell a customer not only how much she spent on Uber last month, but, " 'By the way, that’s twice as much as you’ve spent in the last three months, is there something wrong here?' Something that gets the customer to go hmmm, and think more about their financial health overall.”

Customer service

“ A prediction by Gartner states that over 85% of customer-enterprise relationships will happen without much human interaction by the year 2020. Virtual assistants are all set to become the face of organizations”

OCR

Processing documents

Contract Analyzer

“JP Morgan has harnessed the power of this application of AI, leading to freeing 360,000 hours (yearly) from its employees’ load in only a few seconds.”

Algorithmic Trading

“In 2017, according to Techfunnel, as much as 73% of daily trading activity was carried out by ML algorithms. “

Investment

Read and process public company filings and conference call transcripts.

Sentiment Analysis

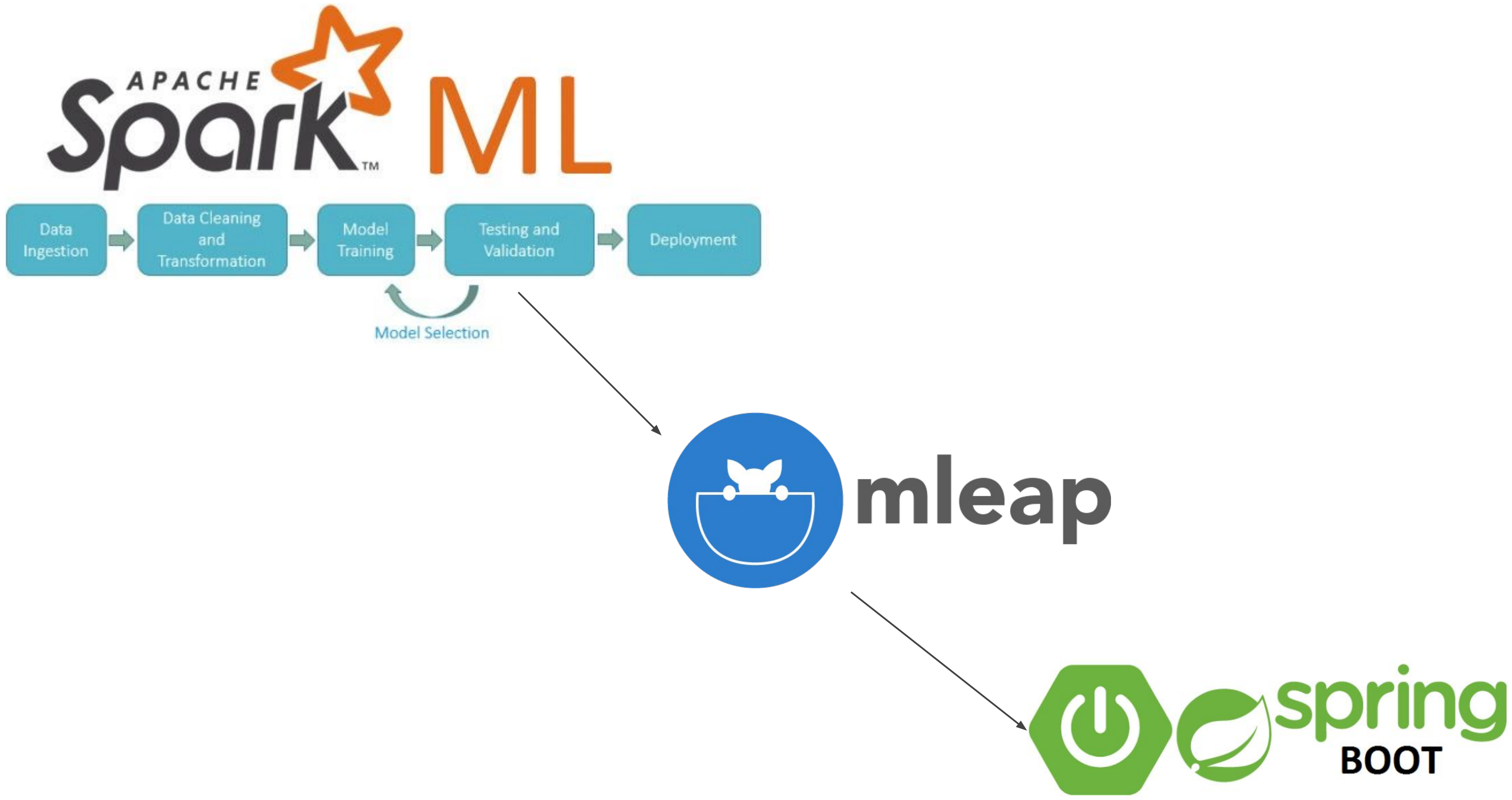
Can be used for due diligences for companies and managers.

Credit scores.

Banca Transilvania use case

Understand a transaction type : Pension, Salary, Something Else

Technologies used: Spark ML, Mleap, Spring-boot, …



APACHE Spark ML

Data Ingestion

Data Cleaning and Transformation

Model Training

Testing and Validation

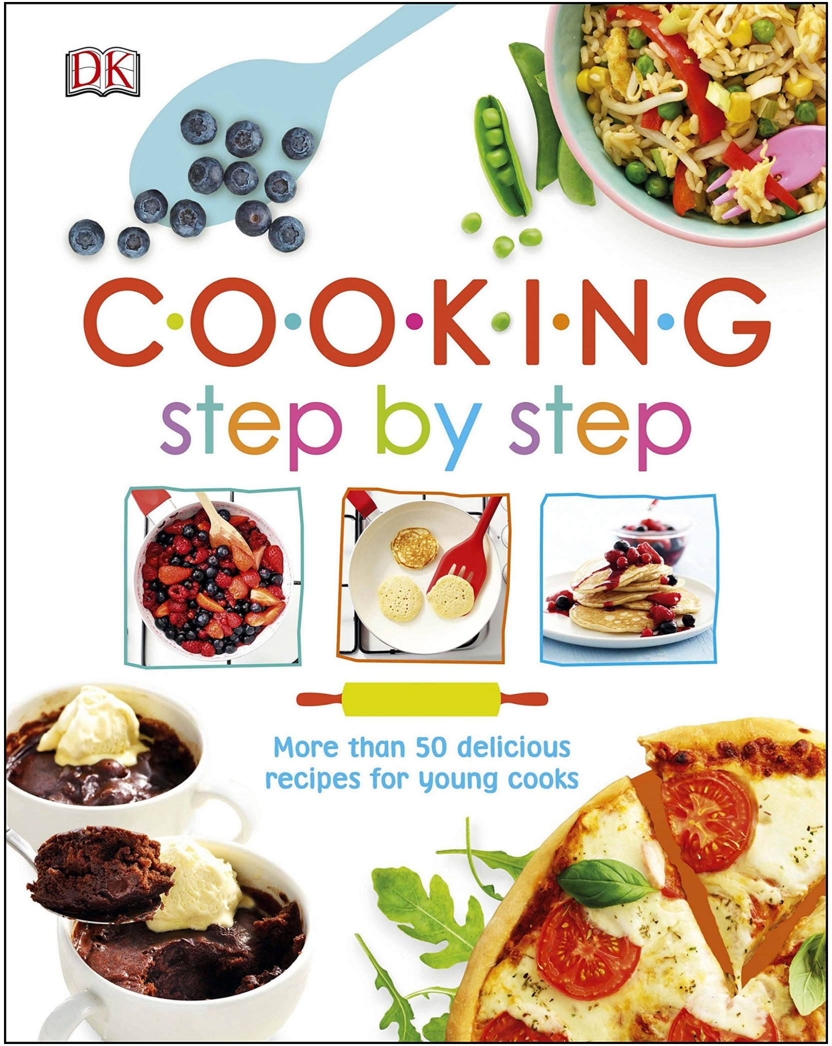
Deployment

Model Selection

mleap

spring BOOT

The usual “Go-to” Technologies



THE RECIPE FOR NLP

1. Gather the data!

- Product reviews
- Products viewed
- Social media posts
- Transactions
- ...

LABELS , LABELS, LABELS!

2. Clean your data!

- Remove all irrelevant characters such as any non alphanumeric characters
- Tokenization
- Lower case vs upper case
- Lemmatization vs stemming
- Stop words removal

Stemming

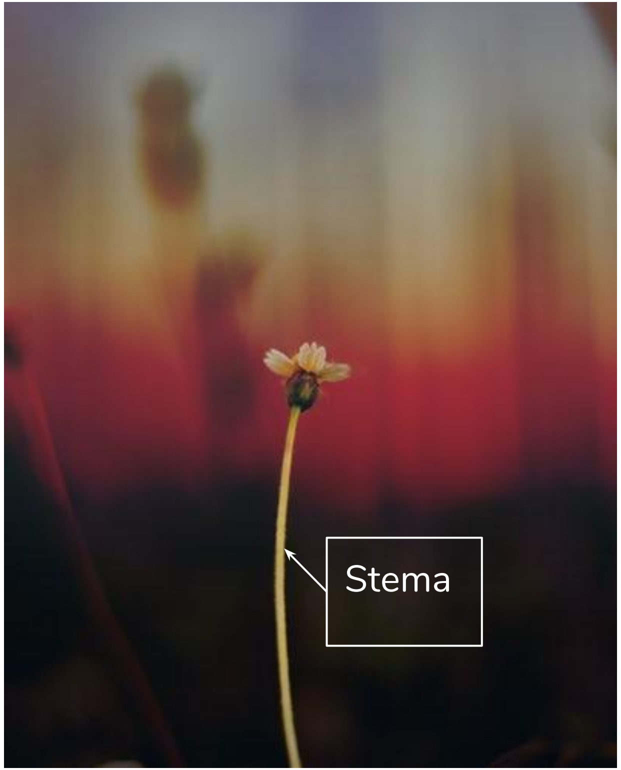
Lookup table or suffix-stripping algorithms

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

Running, runs, runned -> Run

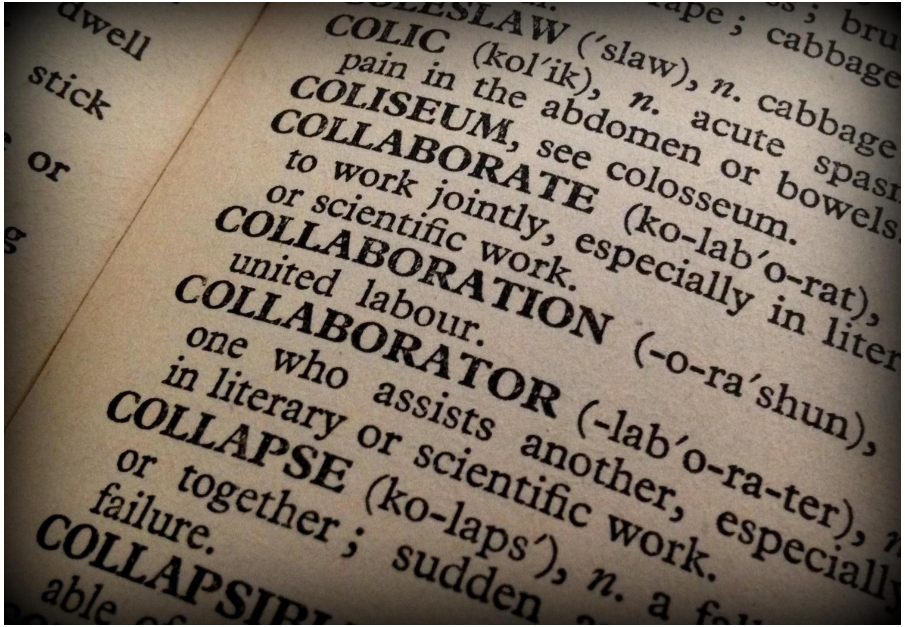
Flower, flowers, flowery, flowering -> Flower

argue, argued, argues, arguing, and argus -> argu.



Lemmatisation

“better” -> “good” lemma
“better” -> “better” after stemming
“saw” -> “see” for verb
“saw” for noun
“saw” -> “saw” after stemming



dwell
stick
or

LESLAW (′slaw), n. cabbage
COLIC (kol′ik), n. acute spasm
pain in the abdomen or bowels
COLISEUM, see colosseum.
COLLABORATE (ko-lab′o-rat),
to work jointly, especially in liter
or scientific work.
COLLABORATION (-o-ra′shun),
united labour.
COLLABORATOR (-lab′o-ra-ter),
one who assists another, especially
in literary or scientific work.
COLLAPSE (ko-laps′), n. a fa
or together; sudden
failure.
COLLAPSIBL
able

Stop words removal

dogs ~~are the~~ best

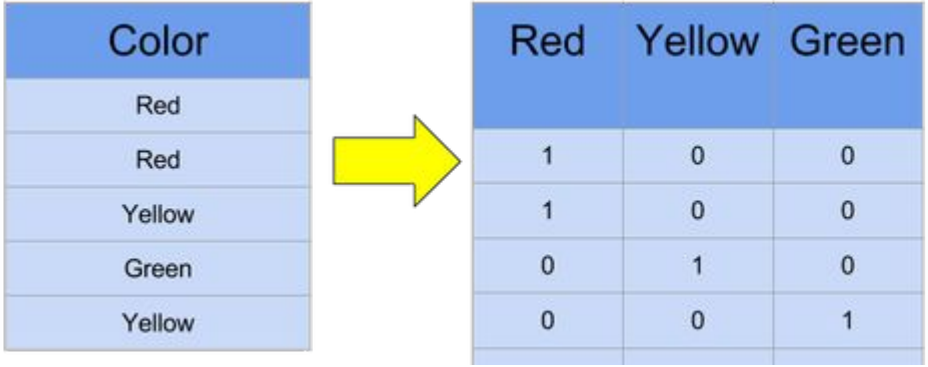
3. Data representation

NUMBERS

NUMBERS

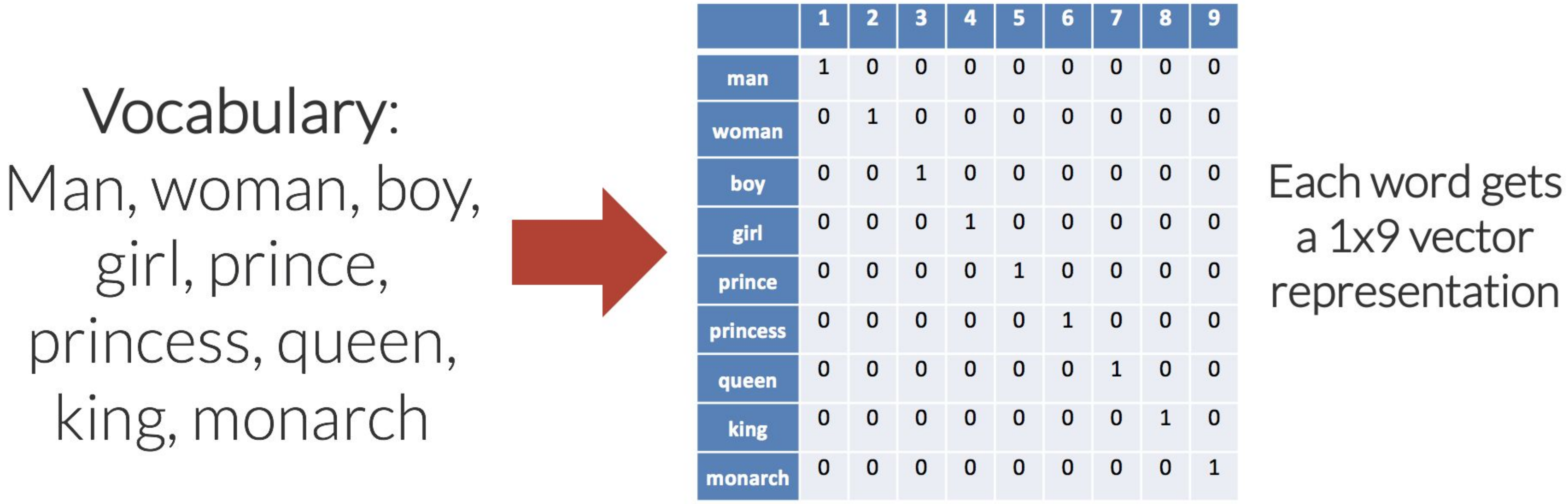
NUMBERS

One-hot encoding (Bag of Words)



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1



Vocabulary:
Man, woman, boy,
girl, prince,
princess, queen,
king, monarch

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

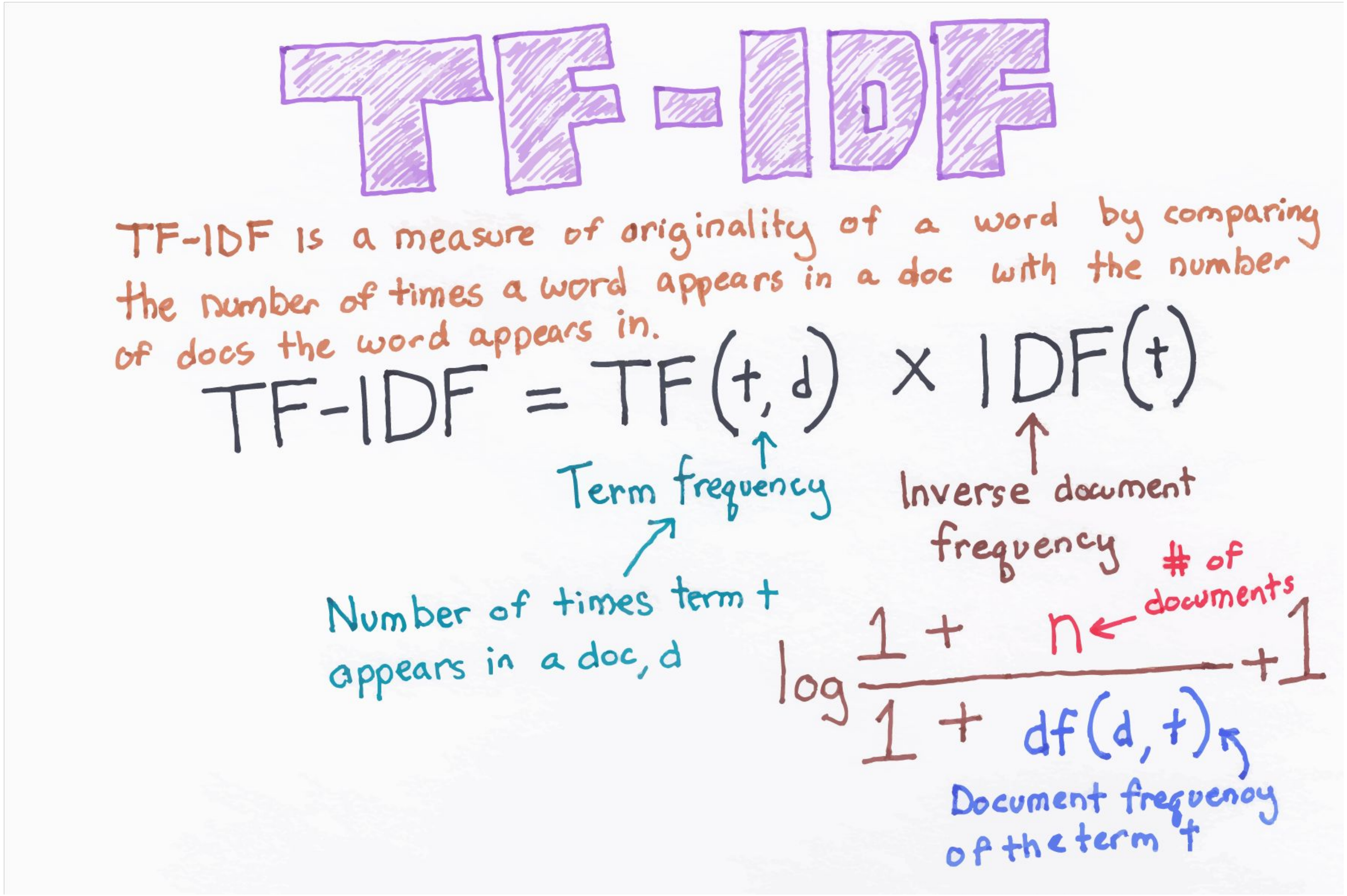
Each word gets
a 1x9 vector
representation

embedding space has the same number of dimensions as the number of words in the vocabulary

CountVectorizer vs TF-IDF

CountVectorizer - Counts the number of occurrences for a word in a **document**

TF-IDF - Term Frequency - Inverted Document Frequency



TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$TF\text{-}IDF = TF(t, d) \times IDF(t)$$

Term frequency

Number of times term t appears in a doc, d

Inverse document frequency

$$\log \frac{1 + n}{1 + df(d, t)} + 1$$

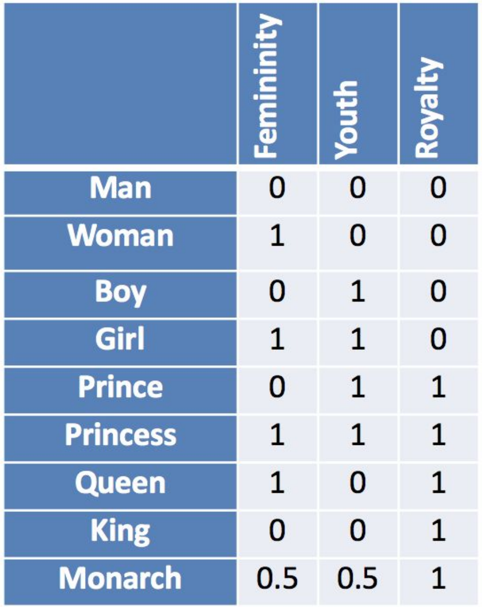
# of documents

Document frequency of the term t

Word Embeddings

Try to build a lower dimensional embedding

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

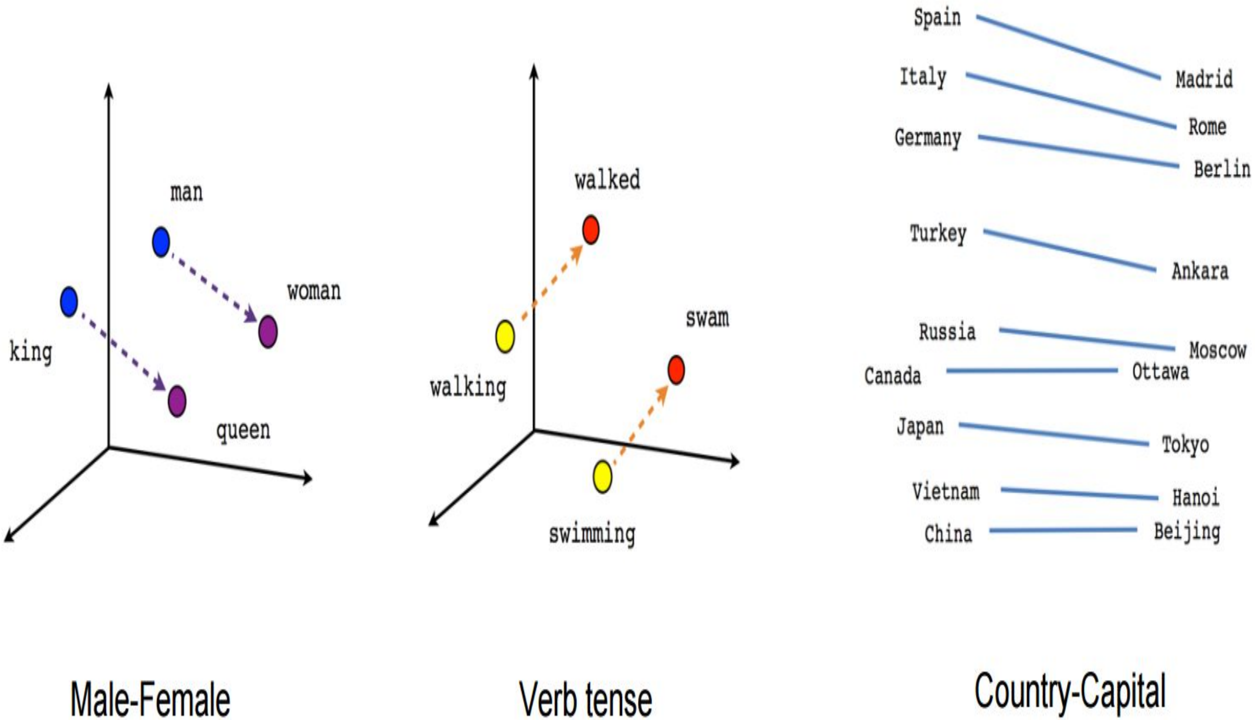


	Femininity	Youth	Royalty
Man	0	0	0
Woman	1	0	0
Boy	0	1	0
Girl	1	1	0
Prince	0	1	1
Princess	1	1	1
Queen	1	0	1
King	0	0	1
Monarch	0.5	0.5	1

@shane_a_lynn | @TeamEdgeTier

Each word gets a 1x3 vector

Similar words… similar vectors



Male-Female

Verb tense

Country-Capital

Word2Vec algorithm from Google

GloVe algorithm from Stanford

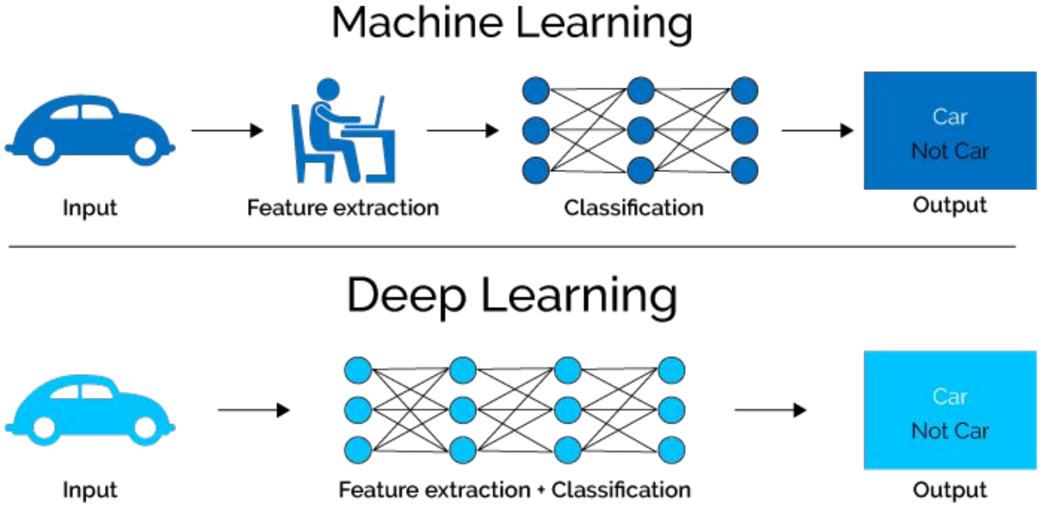
fasttext algorithm from Facebook

GLOVE works similarly as Word2Vec. Word2Vec is a "predictive" model that predicts context given word
GLOVE learns by constructing a co-occurrence matrix (words X context) that basically count how frequently a word appears in a context

4. Machine Learning Model

START SIMPLE

Start Small



Machine Learning

Input → Feature extraction → Classification → Car / Not Car

Output

Deep Learning

Input → Feature extraction + Classification → Car / Not Car

Output

5: Inspection

Confusion Matrix

Accuracy

F-Score

ROC / AUC

Debug String

Understand the errors

THE TECHNICAL PART